Micro-Bench: A Vision-Language Benchmark for Microscopy Understanding

Alejandro Lozano *
Department of Biomedical Data Science
Stanford University
Stanford, CA 94305
lozanoe@stanford.edu

Jeffrey Nirschl *
Department of Pathology
Stanford University
Stanford, CA 94305
jnirschl@stanford.edu

James Burgess
ICME
Stanford University
Stanford, CA 94305
jmhb@stanford.edu

Sanket Rajan Gupte
Department of Computer Science
Stanford University
Stanford, CA 94305
sanketg@stanford.edu

Yuhui Zhang
Department of Computer Science
Stanford University
Stanford, CA 94305
yuhuiz@stanford.edu

Alyssa Unell
Department of Computer Science
Stanford University
Stanford, CA 94305
aunell@stanford.edu

Serena Yeung-Levy
Department of Biomedical Data Science
Stanford University
Stanford, CA 94305
syyeung@stanford.edu

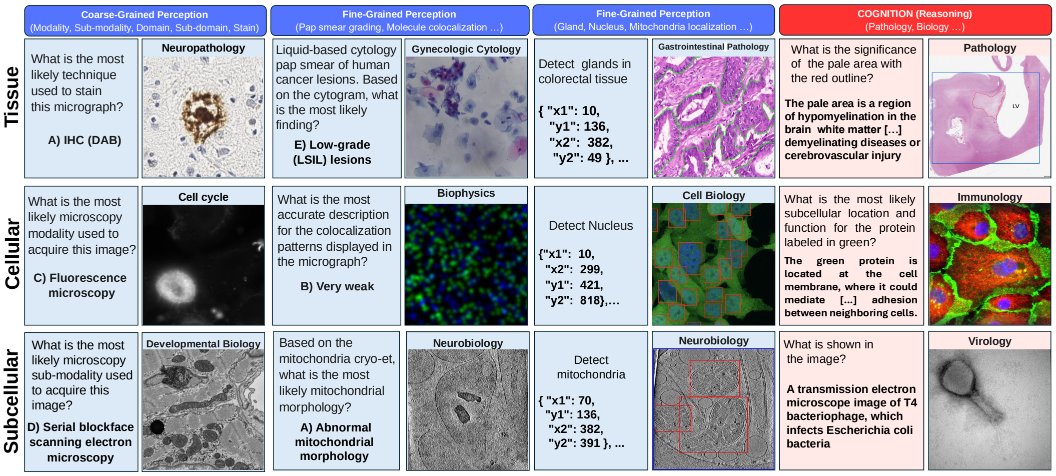
Abstract

Recent advances in microscopy have enabled the rapid generation of terabytes of image data in cell biology and biomedical research. Vision-language models (VLMs) offer a promising solution for large-scale biological image analysis, enhancing researchers' efficiency, identifying new image biomarkers, and accelerating hypothesis generation and scientific discovery. However, there is a lack of standardized, diverse, and large-scale vision-language benchmarks to evaluate VLMs' perception and cognition capabilities in biological image understanding. To address this gap, we introduce Micro-Bench, an expert-curated benchmark encompassing 24 biomedical tasks across various scientific disciplines (biology, pathology), microscopy modalities (electron, fluorescence, light), scales (subcellular, cellular, tissue), and organisms in both normal and abnormal states. We evaluate state-of-the-art biomedical, pathology, and general VLMs on Micro-Bench and find that: i) current models struggle on all categories, even for basic tasks such as distinguishing microscopy modalities; ii) current specialist models fine-tuned on biomedical data often perform worse than generalist models; iii) fine-tuning in specific microscopy domains can cause catastrophic forgetting, eroding prior biomedical knowledge encoded in their base model. iv) weight interpolation between fine-tuned and pretrained models offers one solution to forgetting and improves general performance across biomedical tasks. We release Micro-Bench under a permissive license [2] to accelerate the research and development of microscopy foundation models.

[*] These authors contributed equally.

[2] The dataset is hosted on Hugging Face at: https://huggingface.co/datasets/jnirschl/uBench. We publish full code to replicate results at https://github.com/yeung-lab/Micro-Bench.

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.



What is the most likely technique used to stain this micrograph?

A) IHC (DAB)

Liquid-based cytology pap smear of human cancer lesions. Based on the cytogram, what is the most likely finding?

E) Low-grade (LSIL) lesions

Detect glands in colorectal tissue

{ "x1": 10, "y1": 136, "x2": 382, "y2": 49 }, ...

What is the significance of the pale area with the red outline?

The pale area is a region of hypomyelination in the brain white matter […] demyelinating diseases or cerebrovascular injury

What is the most likely microscopy modality used to acquire this image?

C) Fluorescence microscopy

What is the most accurate description for the colocalization patterns displayed in the micrograph?

B) Very weak

Detect Nucleus

{"x1": 10, "x2": 299, "y1": 421, "y2": 818},…

What is the most likely subcellular location and function for the protein labeled in green?

The green protein is located at the cell membrane, where it could mediate […] adhesion between neighboring cells.

What is the most likely microscopy sub-modality used to acquire this image?

D) Serial blockface scanning electron microscopy

Based on the mitochondria cryo-et, what is the most likely mitochondrial morphology?

A) Abnormal mitochondrial morphology

Detect mitochondria

{ "x1": 70, "y1": 136, "x2": 382, "y2": 391 }, ...

What is shown in the image?

A transmission electron microscope image of T4 bacteriophage, which infects Escherichia coli bacteria

Figure 1: Data samples from Micro-Bench, covering perception (left) and cognition (right) tasks across subcellular, cellular, and tissue levels tasks across electron, fluorescence, and light microscopy.

1 Introduction

Microscopy is a cornerstone of biomedical research [19, 54], enabling detailed study of structures at multiple scales [72]. Advances in cryo-electron microscopy, high-throughput fluorescence microscopy, and whole-slide imaging allow scientists to examine and analyze atomic, subcellular, cellular, and tissue-level structures with high precision [21] to reveal new insights into complex biological processes. To achieve this, researchers interpret and contextualize image findings within their existing scientific knowledge to link observations to biological functions and disease relevance [69]. This process requires domain expertise to identify normal and abnormal states, relate observations to molecular and cellular mechanisms, and distinguish artifacts from meaningful findings. Furthermore, learning the nuances of interpreting images across diverse samples or microscopy modalities — beyond an expert's narrow specialization — involves significant trial and error, creating a bottleneck in analyzing the rapidly increasing volumes of imaging data.

Text is an intuitive interface for interactive analysis, and thus vision-language models (VLMs) are one promising approach to assist with image interpretation. A biomedical VLM incorporating knowledge from diverse microscopy images and insights from multiple domain experts could democratize access to scientific knowledge. Such a model could enable text or chat-guided image analysis, help scientists interpret microscopy images outside their field of expertise, facilitate large-scale image annotation, and connect image findings to relevant literature, genes, small molecule therapeutics, and diseases — potentially accelerating scientific research, hypothesis generation, and discovery [23, 70, 64, 52, 57, 12, 46, 46, 41, 28]. To accomplish this, biomedical VLMs must first accurately interpret microscopy data; recognizing basic features such as modality, stains, sample types, and use this information to reason about an image.

However, there is a lack of diverse, large-scale vision-language benchmarks to evaluate image interpretation across multiple microscopy modalities, scales, organisms, and biological states. Existing benchmarks often focus on single-domain diagnostic capabilities (predominantly encompassing histopathology [27, 31]) rather than understanding and describing the fundamental biological mechanisms driving those outcomes . Moreover, in contrast to current trends in the general vision-language community, current biomedical benchmarks lack comprehensive characterization across both perception (e.g., recognizing, localizing, and counting objects) and cognition capabilities (integrating perceptual attributes and knowledge to deduce more complex answers) [22, 44]. This gap hinders tracking the progress and development of robust VLMs tailored for biomedical research.

To address this gap, we present two contributions:

1. **Micro-Bench**: An expert-curated vision-language benchmark comprising 17,235 microscopy images from a diverse collection of 26 published and unpublished datasets, featuring new annotations and a permissive license. Micro-Bench adopts a holistic approach with

three subsets: two perceptual subsets, including five perception coarse-grained tasks (e.g., domain, modality, and stain identification) and 18 perception fine-grained tasks (e.g., cell type classification and segmentation of mitochondria, nucleolus, and glands), along with a cognition component requiring reasoning about images. In total, Micro-Bench includes 24 tasks across light, fluorescence, and electron microscopy (covering 8 submodalities), 25 staining techniques, and 12 scientific domains. Our benchmark is formatted for both closed visual question answering (VQA) and captioning, enabling the evaluation of generative and embedding models .

2. **Characterization** We leverage Micro-Bench to characterize state-of-the-art general and domain-specific biomedical VLMs. We show several findings: even the best-performing VLMs have high error rates across all microscopy tasks (do not generalize well); specialist biomedical VLMs often underperform general VLMs; specialist fine-tuning in specific domains can cause catastrophic forgetting of biological knowledge that existed in the base model; and a simple weight ensembling strategy can mitigate the forgetting problem within Micro-Bench.

2 Related Work

Benchmarking Vision-Language Models Benchmarks enable the characterization of model behavior, aid the community to better understand AI technology, and influence its development [47]. In the general VLM community, there is an emergence of systematic and holistic model evaluation [79, 44, 78]. To achieve this, VLM benchmark generation typically involves repurposing existing datasets across a collection of multiple tasks [88, 82] and providing a large-scale and standardized evaluation across different tasks [85]. It is usual to organize similar tasks with multiple levels of questions—such as perception and cognition—each further categorized into specific sub-tasks [22]. Once structured, evaluations are typically conducted using one of two approaches: open or VQA. Open VQA can offer more comprehensive insights by simulating the open-ended nature of real-world usage. However, it may require additional experts or models to assess responses, which can introduce potential biases or hallucinations. Alternatively, natural generation metrics may be used, though they often correlate weakly with human responses [20]. Closed VQA mitigates these issues, but requires metadata curated by experts to create meaningful and challenging distractors.

Biomedical Vision-Language Benchmarks. While previous works have developed various biomedical vision-language benchmarks that have been instrumental in advancing diagnostic capabilities, they present three main problems: 1) Task simplicity: Most biomedical computer vision benchmarks predominantly focus on narrow classification and segmentation tasks [11, 62, 5]. 2) Lack of diversity: existing vision-language datasets are usually limited to diagnostic imaging such as radiology or pathology [59, 17]; As shown in Table 12, there is a lack of benchmarks for basic research microscopy. 3) Limited Accessibility: While there are large and diverse datasets encompasisng multiple modalities within microscopy, such as PMC-15M [86], , this data is not yet publicly accessible for training or evaluation.

Vision-Language Models in Biomedicine. Vision-language models (VLMs) can be generally categorized into two types: 1) Contrastive models, such as CLIP [58] and ALIGN [36], which use contrastive learning to create shared image-text embeddings, facilitating tasks like zero-shot classification and text-image retrieval; and 2) Auto-regressive models, such as Flamingo [4] and GPT-4 [3], which integrate image embeddings with large language models (LLMs) to perform zero-shot tasks, follow instructions, and reason about content.

While vision-language models (VLMs) have significant potential to advance biomedicine [55, 69, 23, 70, 50], they are primarily trained on general datasets with limited biomedical coverage, leading to suboptimal performance on biomedical tasks [63, 89]. Hence, specialized VLMs have been developed by fine-tuning generalist models on biomedical data. Notable examples of contrastive models include BiomedCLIP [86], trained on images from PubMed, and histopathology vision-language models such as PLIP [32] and CONCH [83], which were trained on Twitter (now X) and pathology-specific literature, respectively. Despite the advancements in model development, there is still a lack of comprehensive benchmarks to evaluate the performance of image-based perception and reasoning about microscopy images across diverse scales and modalities, limiting our understanding of failure modes within these domains. Our work addresses this issue by providing a comprehensive benchmark across diverse biological processes, organisms, microscopy modalities, domains, and tasks.

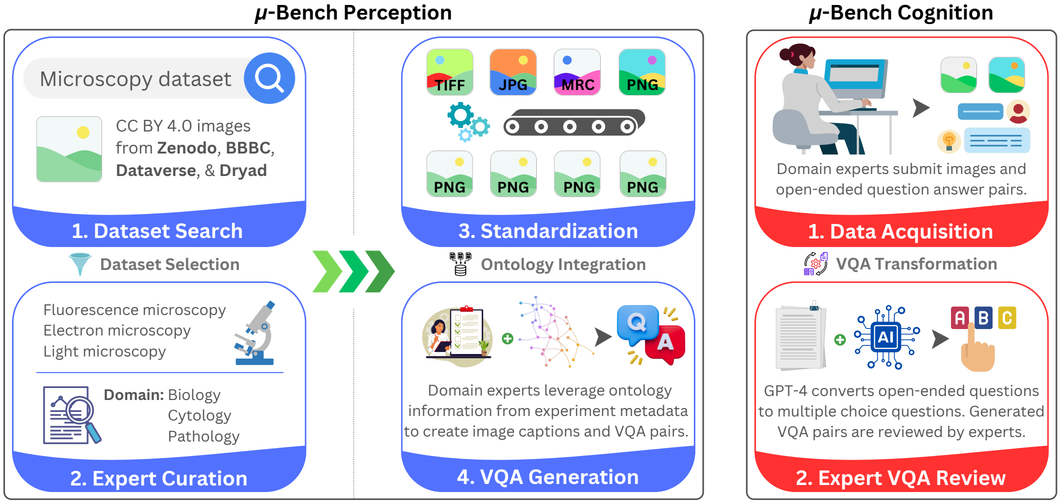

μ-Bench Perception
Microscopy dataset
CC BY 4.0 images from Zenodo, BBBC, Dataverse, & Dryad
1. Dataset Search
Dataset Selection
Fluorescence microscopy
Electron microscopy
Light microscopy
Domain: Biology
Cytology
Pathology
2. Expert Curation
TIFF
JPG
MRC
PNG
PNG
PNG
PNG
PNG
3. Standardization
Ontology Integration
Q
A
Domain experts leverage ontology information from experiment metadata to create image captions and VQA pairs.
4. VQA Generation
μ-Bench Cognition
Domain experts submit images and open-ended question answer pairs.
1. Data Acquisition
VQA Transformation
AI
A
B
C
GPT-4 converts open-ended questions to multiple choice questions. Generated VQA pairs are reviewed by experts.
2. Expert VQA Review

Figure 2: Micro-Bench construction protocol. **Perception dataset (left):** experts taxonomize use cases across subcellular, cellular, and tissue-level applications and collect representative datasets spanning multiple imaging modalities to test those scenarios. Next, datasets are converted to a common format, and the ontological information extracted from their metadata is manually standardized. Aided by this information, experts synthesize VQA pairs designed to test perception ability. **Cognition dataset (right):** First, domain experts use an interactive web application to upload their images and corresponding open-ended VQA pairs. Next, GPT-4 transforms the VQA pairs into a close-ended multiple-choice format. All GPT-4 generations are reviewed by experts before being incorporated into the cognition dataset.

3 Dataset collection methodology

Recognizing the need for an expert-level benchmark in microscopy for comprehensive biological and biomedical understanding, we developed a benchmark to assess VLMs' perception and cognition capabilities in microscopy image analysis following the methodology shown in Figure 2. At a high level, the pipeline consists of two main components: (i) Biomedical experts categorized potential tasks and collected diverse microscopy datasets across multiple scientific domains, focusing on evaluating perception capabilities. (ii) We then complement Micro-Bench by crowdsourcing questions from a larger group of expert microscopists using a web application.

3.1 Perception Dataset Curation

Dataset Review and Selection Open data repositories, including Zenodo, Dataverse, Dryad, and BBBC, among others, were searched for microscopy biomedical image datasets. Data with permissive licenses (CC BY 4.0) allowing derivatives and redistribution were prioritized. A cell biologist and pathologist reviewed the images to ensure high quality (e.g., absence of artifacts or distortion). Diverse datasets were selected to include important biological processes (e.g., cell cycle), organelles (mitochondria, nucleus), and cell/tissue types (e.g., HeLa and HEK cells/cardiac and colorectal tissue). Efforts were made to include diverse biological structures, microscopy modalities, and fields of study. However, basic microscopy research is a broad field, and future work can fill gaps in coverage.

Standardization The original datasets had different organizational structures, file formats, and often very little metadata. Information regarding the scientific discipline (domain), microscopy method, staining, pixel calibration, and the organism was manually determined by expert review or consulting the original publication. The base experimental metadata was supplemented with manual annotation of multiple bio-ontology identifiers (SNOMED, BTO, FMA, LOINC, UBERON, etc.) to connect the image data with rich biology concepts and relationships knowledge graphs in the future. All images were converted into lossless PNG files at their original resolution and stored with metadata in a paired JSON file. An MD5 checksum was computed for the image data, and each image was assigned a 128-bit unique identifier. The image-JSON pairs were converted into an Apache Arrow file for public distribution and ease of use through Hugging Face datasets [43].

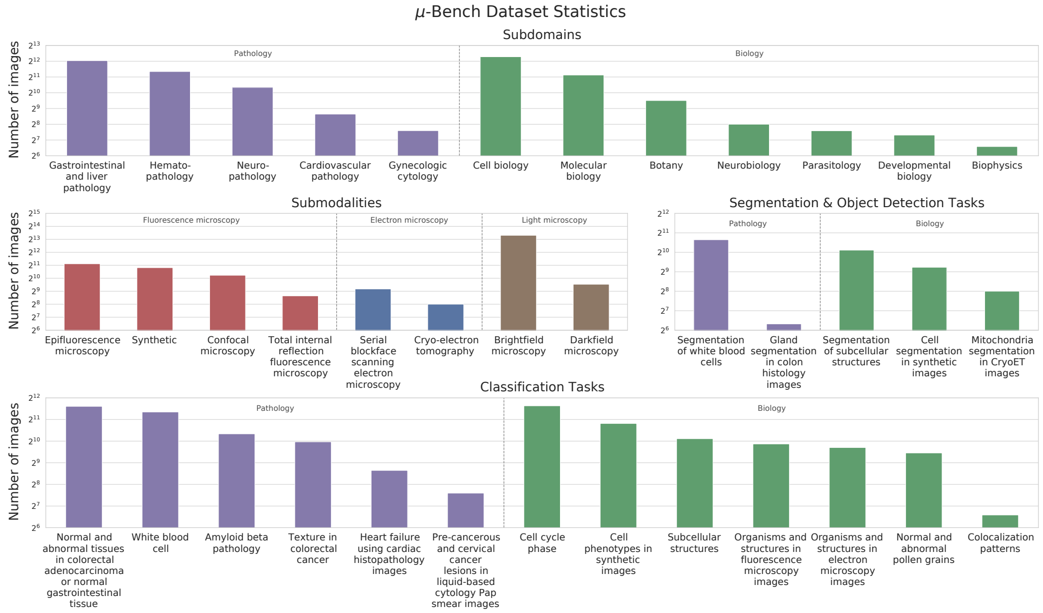


Figure 3: Micro-Bench Perception dataset statistics. The Perception benchmark consists of microscopy images from 12 subdomains in Biology and Pathology, obtained using 8 different imaging techniques, including light, fluorescence, and electron microscopy. It includes 18 perception fine-grained tasks: 13 for classification and 5 for segmentation or object detection.

VQA task generation We used the collected standardized metadata to create closed VQA questions that test capabilities at different levels: easier *coarse-grained* perception and challenging *fine-grained* perception (examples are shown in Figure 1).

The coarse-grained perception split tests basic image properties: the broad category of scientific discipline, the type of microscope, or the stain/contrast agent. These groups are visually distinct (e.g., fluorescence vs. electron microscopy) and relatively straightforward even for non-biologists, but provide a framework for image-based reasoning, including expected normal/abnormal findings for a given sample, artifacts specific to a modality, etc. Although easy for humans, these tasks are essential to assess whether VLMs have an intuitive understanding of microscopy images. Furthermore, VLMs with accurate coarse-grained perception can serve as an independent check for the scientist when the text and image input are discordant.

The fine-grained perception split is more challenging. Within each category of a scientific discipline or microscopy modality, there are image classes or features that need to be recognized to perform image interpretation. Dataset-specific tasks include identifying cell type, subcellular organelles, cell cycle phase, and other visually distinct biological processes that are important for reasoning about biological images. Solving fine-grained perception relies on finer-grained visual features and is more challenging for humans.

We formulate both coarse-grained and fine-grained perception as closed VQA. We chose this over open VQA as it's simpler to analyze and doesn't rely on LLMs for automatic evaluation. To generate VQA options in coarse-grained perception, we designed a tree encompassing microscopy modalities (Figure 15), scientific domains (Figure 16), and staining techniques (Figure 17), which enables sampling fine-grained options within concepts (e.g., selecting IHC(DAB) and IHC(RED) as likely stain options for question regarding light microscopy).

Quality control Throughout all processing, we validate the schema of each data instance to ensure a consistent format and prevent errors before incorporating into Micro-Bench. The schema includes: *modality*: identification of the microscopy modality (BF, EF, or EM); *submodality*: identification of the microscopy sub-modality (e.g., confocal, phase contrast, or scanning electron microscopy); *domain*: determination of the field of study (e.g., cell biology, histology, or pathology); *subdomain*: identification of the sub-field (e.g., cancer research, neurobiology, or infectious diseases); *staining*:

recognition of the staining technique (e.g., H&E, DAPI, or IHC). We compute the perceptual hash of each image for image deduplication–the number of near duplicates or perceptually similar images in Micro-Bench is no different from comparable biomedical benchmarks E.5. Ten external experts not involved in the study evaluated Micro-Bench; the consensus was that an expert with training could perform the VQA tasks, although the difficulty varied E.3. External experts had diverse backgroundsfig:enter-label, significant microscopy experience (median 9.5 yrs), and included post-docs, two board-certified pathologists, and one PI 10. Additional information can be found in the appendix.

Localization task generation We also create a spatial localization benchmark split, which involves predicting the bounding box or segmentation mask for nuclei, mitochondria, cells, and glands(examples are shown in Figure 1). Understanding position and layout enables modeling spatial relationships and context and is fundamental to image understanding. Datasets with segmentation were converted to allow instance segmentation, semantic segmentation, and object detection (bounding box and centroid).

3.2 Cognitive Dataset Curation

While perception datasets evaluate the fundamental capabilities of VLMs for microscopy image analysis, they fall short in assessing their ability to use perception to reason about objects. We curated a cognitive dataset to evaluate more advanced aspects like knowledge and reasoning. The cognitive dataset includes questions related to gene pathways, metabolic pathways, cell signaling and signal transduction, cell physiology and function, protein-protein interactions, cell-cell interactions, unique properties of the cell of origin or cell type in the image, cytoskeleton and cell structure or morphology, and drug mechanisms of action. These categories cover fundamental biological concepts and cellular processes to more deeply evaluate VLMs' knowledge in understanding microscopy images.

Cognition Dataset Collection We began by providing detailed guidelines for question creation to experts (see appendix C.4), which ensured consistency and quality across the dataset. Using an internal chat-like web interface, we asked domain experts to submit questions reflecting their daily research activities. We encouraged a focus on questions that required challenging image-based reasoning, domain expertise, interpretation of experimental results, or hypothesis generation.

In addition to submitting questions, experts provided context regarding experimental details, image acquisition, organisms, treatments, and image descriptions. With this comprehensive information, GPT-4V generated answers to the submitted questions. These answers were subsequently reviewed by experts, who evaluated the accuracy and interpretation of the responses.

Multiple-Choice Question Transformation The collected pairs (image, question, GPT-4V answer, feedback) were transformed into multiple-choice questions using GPT-4. This transformation was guided by a carefully designed prompt (appendix E.2), verified by a cell biologist and a pathologist, to ensure the questions are challenging and reflective of real-world scenarios biomedical researchers face. Each transformed question includes an image, a question, and six candidate choices. One choice is correct, while the other five are distractors generated by GPT-4, where one choice is "None of the above." Domain experts verified the validity of the generated questions and manually corrected a small number of questions. Finally, we ensured that correct answers were uniformly distributed among answer choices A to F.

4 Dataset Description

Perception Dataset Statistics For our perception benchmark, we collected a total of 17,235 microscopy images from 26 distinct public datasets (see Table 5) with permissive licensing, prioritizing open CC-BY licenses. To the best of our knowledge, Micro-Bench Perception is the most diverse microscopy vision-language benchmark, spanning light (LM), fluorescence (FM), and electron microscopy (EM), covering 8 microscopy sub-modalities (see Figure 3), 91 unique cells, tissues, and structures over 25 unique staining techniques (see Figure 6). The perception benchmark subset spans this diversity through closed VQA, object detection, and segmentation.

Cognition Dataset Statistics For our cognition benchmark, we collected 54 microscopy images and 121 questions from experts in the field. Entries were received from 6 users across 5 different institutions. The Micro-Bench Cognition dataset encompasses 3 modalities (fluorescence, electron, light) with 12 sub-modalities, 2 domains (pathology and biology) with 14 sub-domains, and 3 scales (nano, micro, macro), covering a diverse range of topics such as pathology, immunology, and virology. Distributions are shown in Appendix Table 11.

5 VLM benchmarking and results

5.1 Benchmarking approach

Data artifacts like Micro-Bench enable studying model behavior within specialist domains. Since our benchmark covers a wide range of biomedical tasks, we can, for the first time, compare biomedical perception and cognition capabilities across microscopy imagining modalities. In this section, we show the utility of Micro-Bench by reporting empirical findings on a range of VLMs.

Table 1: Macro-average accuracy (with bootstrap confidence interval) for coarse-grained and fine-grained perception and cognition (reasoning) in Micro-Bench .

μ-Bench					
Perception (Coarse-Grained)		Perception (Fine-Grained)		Cognition (Reasoning)	
Model	Accuracy (± CI)	Model	Accuracy (± CI)	Model	Accuracy (± CI)
GPT-4o	62.68 (± 0.35)	GPT-4o	51.73 (± 0.82)	GPT-4o	62.00 (± 9.00)
CogVLM	52.05 (± 0.35)	BiomedCLIP	34.65 (± 0.75)	QwenVLM	41.00 (± 10.00)
QwenVLM	49.85 (± 0.35)	CONCH	33.64 (± 0.72)	CogVLM	41.00 (± 10.00)
BiomedCLIP	47.57 (± 0.34)	ALIGN	31.9 (± 0.72)	OpenCLIP	38.33 (± 8.33)
ALIGN	40.7 (± 0.34)	CLIP	30.09 (± 0.71)	ALIGN	31.00 (± 9.00)
OpenCLIP	36.34 (± 0.33)	OpenCLIP	29.36 (± 0.69)	CLIP	28.00 (± 9.00)
PaliGemma	36.29 (± 0.33)	CogVLM	28.18 (± 0.70)	PaliGemma	25.00 (± 8.00)
CLIP	35.41 (± 0.34)	QuiltNet	27.85 (± 0.69)	BiomedCLIP	25.00 (± 8.00)
PLIP	31.11 (± 0.32)	QwenVLM	27.81 (± 0.70)	CONCH	18.00 (± 7.00)
CONCH	27.84 (± 0.31)	PLIP	25.49 (± 0.68)	Random	17.00 (± 7.00)
QuiltNet	26.58 (± 0.31)	PaliGemma	21.29 (± 0.64)	PLIP	17.00 (± 7.00)
Random	18.34 (± 0.27)	Random	19.13 (± 0.60)	QuiltNet	13.00 (± 6.00)

+ 🟩 General autoregressive VLMs 🟦 General contrastive VLMS 🟥 Pathology contrastive VLMS 🟪 Biomedical contrastive VLMS.

To this end, we first categorized VLMs into two groups: generalist models trained on natural images and language, and 'specialist' models, fine-tuned on biomedical data. Within generalist models, we also distinguish between contrastive and auto-regressive models.

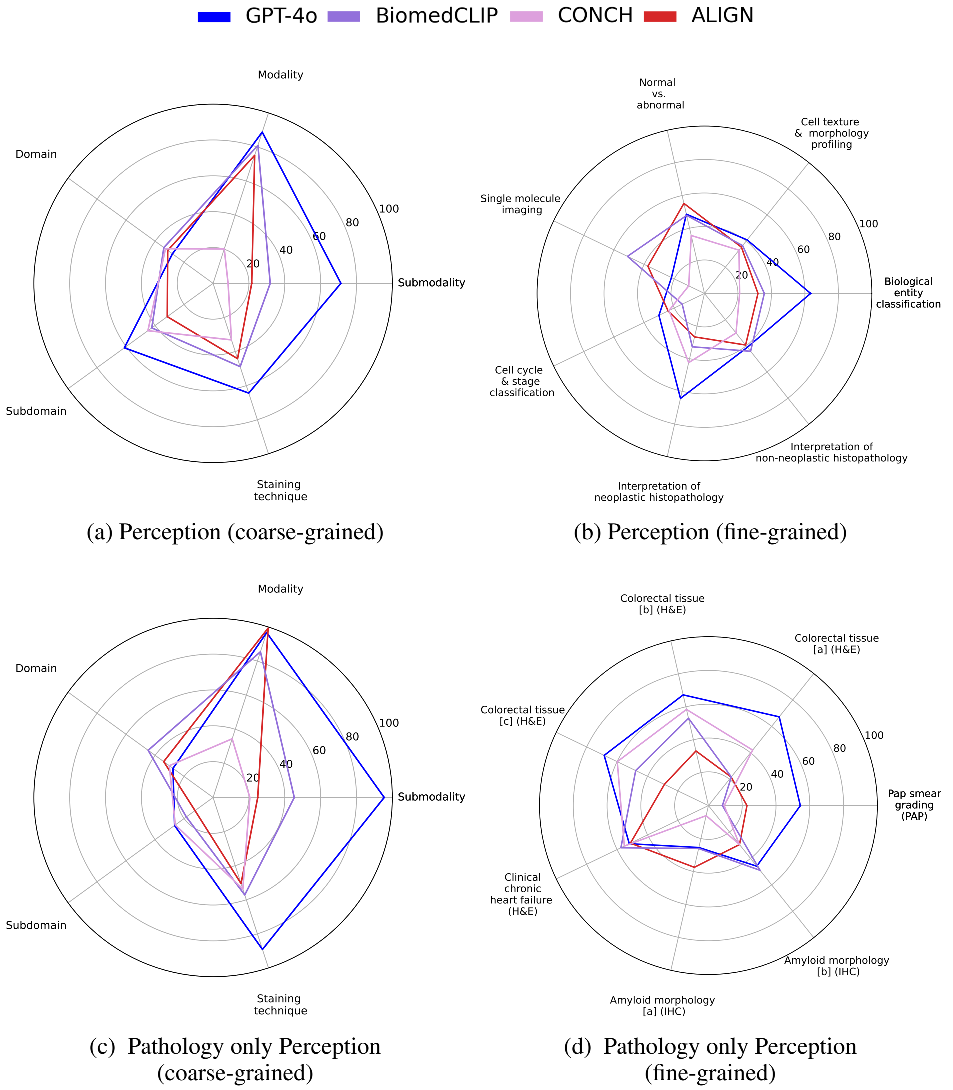
Generalist Contrastive (GC) VLMs We evaluate ALIGN [36], OpenCLIP [14], and CLIP [58] as the canonical contrastive models for natural images. Notably, OpenCLIP and CLIP serve as the initial model weights for numerous finteuned specialist biomedical VLMs.

Generalist autoregressive (GA) VLMs We evaluate with GPT-4o [3], a state-of-the-art enterprise VLM. For open-source models, we test CogVLM [71], QwenVLM [6], and PaliGemma [9] for their strong performance on general domain tasks, instruction-following capabilities. Furthemore, QwenVLM and PaliGemma support object detection.

Specialist contrastive (SC) VLMs Our specialist model selection included two constraints: choosing the best-performing models and preferring minimal architectural changes from their base generalist versions, allowing performance analysis based on variations in training mixtures. We selected BiomedCLIP [86] since it is a robust model with a training set that covers all biomedical imaging modalities in our benchmark (trained on 15 million image-text pairs collected from PubMedCentral). Additionally, we included three pathology VLMs: PLIP (CLIP trained on H&E) [32], QuiltNet (CLIP trained on H&E and IHC) [34], and CONCH (CoCa trained on H&E and IHC) [49], with training dataset sizes of 208k, 1 million, and 1.2 million, respectively. While CONCH and BiomedCLIP are based on CoCa [83] and OpenCLIP, respectively, they modify the architecture or training strategy.

Evaluation The Closed VQA component of Micro-Bench was evaluated with accuracy, generating confidence intervals (CI) via bootstrap [16] (appendix H.2). Object detection was evaluated for models with object detection capabilities (PaliGemma and QwenVLM) in open VQA format using the GRIT localization metric [25] as adopted by prior works [6].

5.2 Results



(a) Perception (coarse-grained)

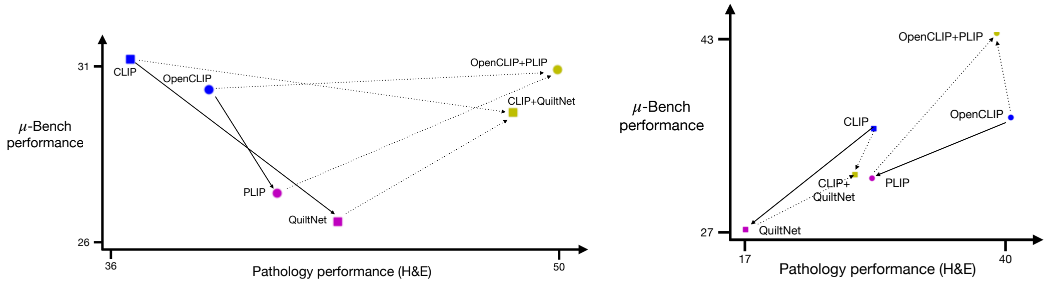
(b) Perception (fine-grained)

(c) Pathology only Perception (coarse-grained)

(d) Pathology only Perception (fine-grained)

Figure 4: Performance comparison on the perception benchmark for the best-performing general domain auto-regressive model (GPT-4o), contrastive model (ALIGN), specialist biomedical contrastive model (BiomedCLIP), and specialist pathology contrastive model (CONCH). The top row shows performance in all of the Micro-Bench while the bottom row shows pathology-only samples.

All models have high error rates Table 1 shows the accuracy across perception and cognition. Even the top-performing model (GPT-4o) has high error rates, with an accuracy of 62.6% on coarse-grained perception, 51.7% on fine-grained perception, and 62.0% on cognition tasks. *On average* GPT-4o also outperforms BiomedCLIP (the best biomedical SC VLM) by a minimum of 15% in all evaluation dimensions and CONCH (the best pathology SC VLM) in pathology-specific perception tasks, showing a difference of 39.37% on coarse-grained and 19.40% in fined-grained tasks (as illustrated in Table 1. However, finer subgroup analysis (Figure 11) shows that GPT-4o does not excel in all perception tasks, including domain identification (coarse-grained), single-molecule imaging,



(a) Perception (fine-grained)

(b) Perception (coarse-grained)

Figure 5: Fine-tuning and microscopy perception generalization on Micro-Bench . Base CLIP models (blue) are fine-tuned to PLIP and QuiltNet using pathology data mixtures (pink). Weight-merging base models with their corresponding fine-tuned models (olive) improves specialist zero-shot performance on Micro-Bench coarse-grained (**Left**) and fined-grained (**Right**) perception.

normal vs abnormal classification, and non-neoplastic histopathology interpretation (fine-grained). The model architecture and training data for GPT-4 are closed source, making it challenging to conclude from these results. However, GPT-4o's high error rates, its substantial gap compared to SC models, and performance variation across task subgroups highlight that μ-Bench is challenging and is not saturated by state-of-the-art general, biomedical, and pathology models.

Specialist biomedical models are often worse than non-specialist models While specialist models are explicitly developed for the biomedical domain, they often underperform non-specialized open-source models. For example, in both coarse-grained perception and cognition tasks (Table 1), GA models (CogVLM and QwenVLM) outperform the best SC model (BiomedCLIP) by 4.4% and 16.0% margins respectively. While GA models have a different training objective, larger training mixture, and more model parameters, a similar trend is observed with GC models (ALIGN, OpenCLIP, and CLIP) as they outperform all pathology VLMs in the same tasks by at least 9.5% (PLIP- ALIGN) and 20.3% (CONCH - OpenCLIP) respectively. This ranking is reversed in fine-grained perception tasks, where BiomedCLIP and CONCH perform best. Indeed, fine-grained perception closely resembles the data mixture used to fine-tune contrastive specialist models [31]. This characterization shows weakness in current microscopy biomedical model development, which we investigate next.

Specialist training can cause catastrophic forgetting Generalist contrastive models like (OpenCLIP and CLIP) surprisingly outperform their fine-tuned counterparts (PILP and QuiltNet) in coarse-grained perception and cognition (Table 1). Specifically, PILP and QuiltNet are fine-tuned directly from OpenCLIP and CLIP using pathology-only (brightfield microscopy) data closest to Micro-Bench fine-grained perception tasks. Although it improves performance on pathology-specific perception fine-grained tasks (Figure 4), it degrades performance on all other tasks (Table 1).

Micro-Bench characterization drives robust model development To address catastrophic forgetting identified in our multi-level evaluation, we ensemble base model weights (OpenCLIP / CLIP) with fine-tuned model weights (PLIP/QuiltNet) to create merged models (PLIP+OpenCLIP / QuiltNet+CLIP), as proposed in [75]. As shown in Figure 5, when comparing merged models to their fine-tuned counterparts, perception performance increases across all of Micro-Bench (y-axis), including pathology-specific tasks (x-axis). This is remarkable given that model merging (or model ensembling) is a simple and training-free strategy. To our knowledge, this is the first application of merging to biomedical embedding models, suggesting many further opportunities to create generalizable biomedical models.

Micro-Bench supports probing design decisions for biomedical VLMs We have shown that Micro-Bench offers valuable insights into microscopy biomedical VLMs and hope it encourages further evaluations of design choices. Data diversity is one factor: Table 1 illustrates that BiomedCLIP, trained across all microscopy modalities in Micro-Bench, surpasses specialist models, albeit with a smaller margin for fine-grained tasks compared to CONCH, which uses pathology data. Regarding model architecture and training strategy, generalist autoregressive models (CogVLM and QwenVLM) outperform contrastive models (ALIGN and CLIP) in coarse-grained perception, but the opposite is true for fine-grained perception. For object localization, PaliGemma outperformed QwenVLM (Table

18) on Micro-Bench, though both performed poorly, and no specialist models support detection. Future research could explore prompting strategies, data curation, and new methods to mitigate catastrophic forgetting. One clear opportunity is to fine-tune a base model on multiple tasks or datasets, and merge them all for more a more generalized robust model [74, 61, 24, 60]. A second direction is alternative weight merging strategies [53, 90, 80].

6 Conclusion

Benchmarks drive advancements in machine learning by providing a standard to measure progress and allowing researchers to identify weaknesses in current approaches. Thus, the lack of biomedical vision-language benchmarks limits the ability to develop and evaluate specialist VLMs. We address this gap in microscopy by introducing the most diverse collection of microscopy vision-language tasks spanning perception and cognition. We use Micro-Bench to establish, for the first time, the performance of some of the most capable VLMs available and find high error rates of 30%, highlighting room for improvement. We demonstrate how Micro-Bench can be leveraged to generate new insights. Lastly, we share Micro-Bench to enable researchers to measure progress in microscopy foundation models.

References

[1] Asma Ben Abacha, Sadid A Hasan, Vivek V Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. *CLEF (working notes)*, 2(6), 2019.

[2] Andrea Acevedo, Anna Merino, Santiago Alférez, Ángel Molina, Laura Boldú, and José Rodellar. A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. *Data Brief*, 30(105474):105474, June 2020.

[3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[4] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[5] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.

[6] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[7] Sebastiano Battiato, Alessandro Ortis, Francesca Trenta, Lorenzo Ascari, Mara Politi, and Consolata Siniscalco. Pollen13k: A large scale microscope pollen grain image dataset. *2020 IEEE International Conference on Image Processing (ICIP)*, pages 2456–2460, 2020.

[8] Asma Ben Abacha, Mourad Sarrouti, Dina Demner-Fushman, Sadid A Hasan, and Henning Müller. Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain. In *Proceedings of the CLEF 2021 Conference and Labs of the Evaluation Forum-working notes*. 21-24 September 2021, 2021.

[9] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. PaliGemma: A versatile 3B VLM for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

[10] James Burgess, Jeffrey J Nirschl, Maria-Clara Zanellati, Alejandro Lozano, Sarah Cohen, and Serena Yeung-Levy. Orientation-invariant autoencoders learn robust representations for shape profiling of cells and organelles. *Nat. Commun.*, 15(1), February 2024.

[11] Juan C Caicedo, Allen Goodman, Kyle W Karhohs, Beth A Cimini, Jeanelle Ackerman, Marzieh Haghighi, CherKeng Heng, Tim Becker, Minh Doan, Claire McQuin, et al. Nucleus segmentation across imaging experiments: the 2018 data science bowl. *Nature methods*, 16(12):1247–1253, 2019.

[12] Anne E Carpenter, Beth A Cimini, and Kevin W Eliceiri. Smart microscopes of the future. *Nature methods*, 20(7):962–964, 2023.

[13] Souradip Chakraborty, Ekaba Bisong, Shweta Bhatt, Thomas Wagner, Riley Elliott, and Francesco Mosconi. Biomedbert: A pre-trained biomedical language model for qa and ir. In *Proceedings of the 28th international conference on computational linguistics*, pages 669–679, 2020.

[14] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023.

[15] Nathan H Cho, Keith C Cheveralls, Andreas-David Brunner, Kibeom Kim, André C Michaelis, Preethi Raghavan, Hirofumi Kobayashi, Laura Savy, Jason Y Li, Hera Canaj, James Y S Kim, Edna M Stewart, Christian Gnann, Frank McCarthy, Joana P Cabrera, Rachel M Brunetti, Bryant B Chhun, Greg Dingle, Marco Y Hein, Bo Huang, Shalin B Mehta, Jonathan S Weissman, Rafael Gómez-Sjöberg, Daniel N Itzhak, Loïc A Royer, Matthias Mann, and Manuel D Leonetti. OpenCell: Endogenous tagging for the cartography of human cellular organization. *Science*, 375(6585):eabi6983, March 2022.

[16] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. Chapman and Hall/CRC, 1994.

[17] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

[18] Philipp Eulenberg, Niklas Köhler, Thomas Blasi, Andrew Filby, Anne E Carpenter, Paul Rees, Fabian J Theis, and F Alexander Wolf. Reconstructing cell cycle and disease progression using deep learning. *Nature communications*, 8(1):463, 2017.

[19] James G Evans and Paul Matsudaira. Linking microscopy and high content screening in large-scale biomedical research. *High Content Screening: A Powerful Approach to Systems Cell Biology and Drug Discovery*, pages 33–38, 2006.

[20] Scott L Fleming, Alejandro Lozano, William J Haberkorn, Jenelle A Jindal, Eduardo P Reis, Rahul Thapa, Louis Blankemeier, Julian Z Genkins, Ethan Steinberg, Ashwin Nayak, et al. Medalign: A clinician-generated dataset for instruction following with electronic medical records. *arXiv preprint arXiv:2308.14089*, 2023.

[21] Laurence Foss. *The end of modern medicine: Biomedical science under a microscope*. SUNY Press, 2002.

[22] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

[23] Shanghua Gao, Ada Fang, Yepeng Huang, Valentina Giunchiglia, Ayush Noori, Jonathan Richard Schwarz, Yasha Ektefaie, Jovana Kondic, and Marinka Zitnik. Empowering biomedical discovery with ai agents. *arXiv preprint arXiv:2404.02831*, 2024.

[24] Almog Gueta, Elad Venezian, Colin Raffel, Noam Slonim, Yoav Katz, and Leshem Choshen. Knowledge is a region in weight space for fine-tuned language models. *arXiv preprint arXiv:2302.04863*, 2023.

[25] Tanmay Gupta, Ryan Marten, Aniruddha Kembhavi, and Derek Hoiem. Grit: General robust image task benchmark. *arXiv preprint arXiv:2204.13653*, 2022.

[26] Sadid A Hasan, Yuan Ling, Oladimeji Farri, Joey Liu, Henning Müller, and Matthew P Lungren. Overview of imageclef 2018 medical domain visual question answering task. In *CLEF (Working Notes)*, 2018.

[27] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

[28] Marco Y Hein, Duo Peng, Verina Todorova, Frank McCarthy, Kibeom Kim, Chad Liu, Laura Savy, Camille Januel, Rodrigo Baltazar-Nunez, Sophie Bax, et al. Global organelle profiling reveals subcellular localization and remodeling at proteome scale. *bioRxiv*, pages 2023–12, 2023.

[29] Michael Held, Michael H A Schmitz, Bernd Fischer, Thomas Walter, Beate Neumann, Michael H Olma, Matthias Peter, Jan Ellenberg, and Daniel W Gerlich. CellCognition: time-resolved phenotype annotation in high-throughput live cell imaging. *Nat. Methods*, 7(9):747–754, September 2010.

[30] Yefan Huang, Xiaoli Wang, Feiyan Liu, and Guofeng Huang. Ovqa: A clinically generated visual question answering dataset. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2924–2938, 2022.

[31] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9):2307–2316, 2023.

[32] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9):2307–2316, 2023.

[33] Elima Hussain, Lipi B Mahanta, Himakshi Borah, and Chandana Ray Das. Liquid based-cytology pap smear dataset for automated multi-class diagnosis of pre-cancerous and cervical cancer lesions. *Data Brief*, 30(105589):105589, June 2020.

[34] Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1m: One million image-text pairs for histopathology. *Advances in Neural Information Processing Systems*, 36, 2024.

[35] Andrii Iudin, Paul K Korir, Sriram Somasundharam, Simone Weyand, Cesare Cattavitello, Neli Fonseca, Osman Salih, Gerard J Kleywegt, and Ardan Patwardhan. Empiar: the electron microscopy public image archive. *Nucleic Acids Research*, 51(D1):D1503–D1511, 2023.

[36] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[37] Changhun Jung, Mohammed Abuhamad, David Mohaisen, Kyungja Han, and DaeHun Nyang. Wbc image classification and generative models based on convolutional neural network. *BMC Medical Imaging*, 22(1):94, 2022.

[38] Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A Valous, Dyke Ferber, Lina Jansen, Constantino Carlos Reyes-Aldasoro, Inka Zörnig, Dirk Jäger, Hermann Brenner, Jenny Chang-Claude, Michael Hoffmeister, and Niels Halama. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS Med.*, 16(1):e1002730, January 2019.

[39] Jakob Nikolas Kather, Cleo-Aron Weis, Francesco Bianconi, Susanne M Melchers, Lothar R Schad, Timo Gaiser, Alexander Marx, and Frank Gerrit Zöllner. Multi-class texture analysis in colorectal cancer histology. *Sci. Rep.*, 6:27988, June 2016.

[40] Olga Kovaleva, Chaitanya Shivade, Satyananda Kashyap, Karina Kanjaria, Joy Wu, Deddeh Ballah, Adam Coy, Alexandros Karargyris, Yufan Guo, David Beymer Beymer, et al. Towards visual dialog for radiology. In *Proceedings of the 19th SIGBioMed workshop on biomedical language processing*, pages 60–69, 2020.

[41] Talley Lambert and Jennifer Waters. Towards effective adoption of novel image analysis methods. *Nature Methods*, 20(7):971–972, 2023.

[42] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.

[43] Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[44] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

[45] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[46] Xinyang Li, Yuanlong Zhang, Jiamin Wu, and Qionghai Dai. Challenges and opportunities in bioimage analysis. *Nature Methods*, 20(7):958–961, 2023.

[47] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

[48] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE, 2021.

[49] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Andrew Zhang, Long Phi Le, et al. Towards a visual-language foundation model for computational pathology. *arXiv preprint arXiv:2307.12914*, 2023.

[50] Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, pages 1–11, 2024.

[51] Jun Ma, Ronald Xie, Shamini Ayyadhury, Cheng Ge, Anubha Gupta, Ritu Gupta, Song Gu, Yao Zhang, Gihun Lee, Joonkee Kim, et al. The multimodality cell segmentation challenge: toward universal solutions. *Nature methods*, pages 1–11, 2024.

[52] Leonel Malacrida. Phasor plots and the future of spectral and lifetime imaging. *Nature Methods*, 20(7):965–967, 2023.

[53] Michael S Matena and Colin A Raffel. Merging models with fisher-weighted averaging. *Advances in Neural Information Processing Systems*, 35:17703–17716, 2022.

[54] Arno P Merkle and Jeff Gelb. The ascent of 3d x-ray microscopy in the laboratory. *Microscopy Today*, 21(2):10–15, 2013.

[55] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

[56] Jeffrey J Nirschl, Andrew Janowczyk, Eliot G Peyster, Renee Frank, Kenneth B Margulies, Michael D Feldman, and Anant Madabhushi. A deep-learning classifier identifies patients with clinical heart failure using whole-slide images of h&e tissue. *PLoS One*, 13(4):e0192726, April 2018.

[57] Damian Dalle Nogare, Matthew Hartley, Joran Deschamps, Jan Ellenberg, and Florian Jug. Using ai in bioimage analysis to elevate the rate of scientific discovery as a community. *Nature methods*, 20(7):973–975, 2023.

[58] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[59] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[60] Alexandre Ramé, Kartik Ahuja, Jianyu Zhang, Matthieu Cord, Léon Bottou, and David Lopez-Paz. Model ratatouille: Recycling diverse models for out-of-distribution generalization. In *International Conference on Machine Learning*, pages 28656–28679. PMLR, 2023.

[61] Alexandre Rame, Matthieu Kirchmeyer, Thibaud Rahier, Alain Rakotomamonjy, Patrick Gallinari, and Matthieu Cord. Diverse weight averaging for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 35:10821–10836, 2022.

[62] Joel Saltz, Rajarsi Gupta, Le Hou, Tahsin Kurc, Pankaj Singh, Vu Nguyen, Dimitris Samaras, Kenneth R Shroyer, Tianhao Zhao, Rebecca Batiste, et al. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell reports*, 23(1):181–193, 2018.

[63] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[64] Morgan Schwartz, Uriah Israel, Xuefei Wang, Emily Laubscher, Changhua Yu, Rohit Dilip, Qilin Li, Joud Mari, Johnathon Soro, Kevin Yu, et al. Scaling biological discovery at the interface of deep learning and cellular imaging. *Nature Methods*, 20(7):956–957, 2023.

[65] Korsuk Sirinukunwattana, Josien P. W. Pluim, Hao Chen, Xiaojuan Qi, PhengAnn Heng, Yun Bo Guo, Li Yang Wang, Bogdan J. Matuszewski, Elia Bruni, Urko Sanchez, Anton B¨ohm, Olaf Ronneberger, Bassem Ben Cheikh, Daniel Racoceanu, Philipp Kainz, Michael Pfeiffer, Martin Urschler, David R. J. Snead, and Nasir M. Rajpoot. Gland segmentation in colon histology images: The glas challenge contest, 2016.

[66] Noor Mohamed Sheerin Sitara and Kavitha Srinivasan. Ssn mlrg at vqa-med 2021: An approach for vqa to solve abnormality related queries using improved datasets. In *CLEF (working notes)*, pages 1329–1335, 2021.

[67] Ziqi Tang, Kangway V Chuang, Charles DeCarli, Lee-Way Jin, Laurel Beckett, Michael J Keiser, and Brittany N Dugger. Interpretable classification of alzheimer's disease pathologies with a convolutional neural network pipeline. *Nat. Commun.*, 10(1):2173, May 2019.

[68] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

[69] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *NEJM AI*, 1(3):AIoa2300138, 2024.

[70] Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

[71] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023.

[72] Michael Weber and Jan Huisken. Multidisciplinarity is critical to unlock the full potential of modern light microscopy. *Frontiers in Cell and Developmental Biology*, 9:739015, 2021.

[73] Daniel R Wong, Ziqi Tang, Nicholas C Mew, Sakshi Das, Justin Athey, Kirsty E McAleese, Julia K Kofler, Margaret E Flanagan, Ewa Borys, Charles L White, 3rd, Atul J Butte, Brittany N Dugger, and Michael J Keiser. Deep learning from multiple experts improves identification of amyloid neuropathologies. *Acta Neuropathol. Commun.*, 10(1):66, April 2022.

[74] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pages 23965–23998. PMLR, 2022.

[75] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7959–7971, 2022.

[76] Gong-Her Wu, Charlene Smith-Geater, Jesús G Galaz-Montoya, Yingli Gu, Sanket R Gupte, Ranen Aviner, Patrick G Mitchell, Joy Hsu, Ricardo Miramontes, Keona Q Wang, Nicolette R Geller, Cathy Hou, Cristina Danita, Lydia-Marie Joubert, Michael F Schmid, Serena Yeung, Judith Frydman, William Mobley, Chengbiao Wu, Leslie M Thompson, and Wah Chiu. CryoET reveals organelle phenotypes in huntington disease patient iPSC-derived and mouse primary neurons. *Nat. Commun.*, 14(1):692, February 2023.

[77] Yong Wu, Mansoureh Eghbali, Jimmy Ou, Rong Lu, Ligia Toro, and Enrico Stefani. Quantitative determination of spatial protein-protein correlations in fluorescence confocal microscopy. *Biophys. J.*, 98(3):493–504, February 2010.

[78] Peng Xia, Ze Chen, Juanxi Tian, Yangrui Gong, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, et al. Cares: A comprehensive benchmark of trustworthiness in medical vision language models. *arXiv preprint arXiv:2406.06007*, 2024.

[79] Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265*, 2023.

[80] Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, 36, 2024.

[81] Jiancheng Yang, Rui Shi, and Bingbing Ni. Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 191–195. IEEE, 2021.

[82] Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Xiaoshui Huang, Zhiyong Wang, Lu Sheng, Lei Bai, et al. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *Advances in Neural Information Processing Systems*, 36, 2024.

[83] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

[84] Longjiang Yu and Shenghe Sun. Image robust hashing based on dct sign. In *2006 International Conference on Intelligent Information Hiding and Multimedia*, pages 131–134. IEEE, 2006.

[85] Duzhen Zhang, Yahan Yu, Chenxing Li, Jiahua Dong, Dan Su, Chenhui Chu, and Dong Yu. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*, 2024.

[86] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

[87] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.

[88] Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. Why are visually-grounded language models bad at image classification? *arXiv preprint arXiv:2405.18415*, 2024.

[89] Zihao Zhao, Yuxiao Liu, Han Wu, Yonghao Li, Sheng Wang, Lin Teng, Disheng Liu, Xiang Li, Zhiming Cui, Qian Wang, et al. Clip in medical imaging: A comprehensive survey. *arXiv preprint arXiv:2312.07353*, 2023.

[90] Max Zimmer, Christoph Spiegel, and Sebastian Pokutta. Sparse model soups: A recipe for improved pruning via model averaging. *arXiv preprint arXiv:2306.16788*, 2023.